Computers in Human Behavior 70 (2017) 207–221

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation

Rod D. Roscoe [a, *], Joshua Wilson [b], Adam C. Johnson [a], Christopher R. Mayra [a]

[a] 7271 E. Sonoran Arroyo Mall, Santa Catalina Hall, Suite 150, Arizona State University-Polytechnic, Mesa, AZ 85212, United States
[b] 213E Willard Hall, School of Education, University of Delaware, Newark, DE 19716, United States

ARTICLE INFO

Article history:
Received 1 June 2016
Received in revised form
12 December 2016
Accepted 28 December 2016
Available online 2 January 2017

Keywords:
Automated writing evaluation
Formative feedback
Technology adoption
User experience
Human-computer interaction
Writing

ABSTRACT

Automated writing evaluation (AWE) is a popular form of educational technology designed to supplement writing instruction and feedback, yet research on the effectiveness of AWE has observed mixed findings. The current study considered how students' perceptions of automated essay scoring and feedback influenced their writing performance, revising behaviors, and future intentions toward the technology. The manner in which the software was presented—claims about the accuracy and quality of the automated scoring and feedback—were modestly related to students' expectations and perceptions. However, students' direct experiences with the software were most strongly associated with their perceptions. Importantly, students' perceptions seemed to have minimal impact on their "in the moment" use of the software to write and revise successfully. Students revised and improved their essays regardless of their positive or negative views of the system. However, positive and negative perceptions significantly predicted future intentions to use the software again or to recommend the software to a friend. Implications for AWE design, implementation, and evaluation are discussed.

1. Introduction

Automated writing evaluation (AWE) is a popular educational technology that saves teachers time in assessing writing, enables more writing practice, and supplements writing instruction. Commercially available systems have been deployed across thousands of classrooms, such as Educational Testing Service's *Criterion* (Burstein, Tetreault, & Madnani, 2013), Pearson's *WriteToLearn* (Foltz, Streeter, Lochbaum, & Landauer, 2013), and Measurement Incorporated's *Project Essay Grade* (PEG, Wilson, Olinghouse, & Andrada, 2014). Other systems serve as test beds for research on writing and AWE design, such as *Writing Pal* (Allen, Crossley, Snow, & McNamara, 2014; Roscoe & McNamara, 2013).

Each technology employs different algorithms but the underlying premises are similar (Dikli, 2006; Shermis & Burstein, 2013). Natural language processing (NLP) tools extract linguistic, structural, semantic, and rhetorical text features related to text quality, and these relationships can be statistically modeled to assign holistic writing scores and assess writing traits. Many systems exhibit high scoring accuracy (Bridgeman, Trapani, & Attali, 2012; Shermis, 2014) and increasingly offer constructive, formative feedback on qualities such as usage, mechanics, organization, and development (e.g., Burstein et al., 2013). These scoring and feedback functions, along with the ability to process thousands of essays in seconds, can enable teachers to offer more writing assignments without a concomitant increase in workload.

Research on the *effectiveness* of AWE—the extent to which students improve in writing proficiency after using the software—has produced mixed findings (Stevenson & Phakiti, 2013). In one experimental evaluation of *Criterion*, Kellogg, Whiteford, and Quinlan (2010) asked freshman composition students to write and revise three essays with varying levels of feedback (i.e., no feedback versus feedback on one or more essay). Although students made fewer mechanical errors, overall writing quality was not affected by the amount of feedback. Other evaluations have examined patterns of revising and proficiency in larger datasets. Wilson et al. (2014) examined the performance of over 4000 students who used PEG to write and revise essays with feedback. Less than a quarter of students (n = 955) submitted more than one revision of their first drafts. Among those who did revise, students achieved small score increases with each draft, but the rate of growth decreased over time and reached a plateau around the 11th

* Corresponding author.
E-mail addresses: rod.roscoe@asu.edu (R.D. Roscoe), joshwils@udel.edu (J. Wilson), acjohn17@asu.edu (A.C. Johnson), cmayra@asu.edu (C.R. Mayra).

http://dx.doi.org/10.1016/j.chb.2016.12.076
0747-5632/

or 12th revision. In their review, Stevenson and Phakiti (2013) concluded that "there is only modest evidence that AWE feedback has a positive effect on the quality of texts that students produce using AWE, and that as yet there is little clarity about whether AWE is associated with more general improvements in writing proficiency" (p. 62).

One cause of these mixed efficacy findings may be similarly mixed beliefs about the appropriateness of automated scoring and feedback. Concerns about validity have been at the heart of long-standing debates about AWE (Anson et al., 2013; Condon, 2013; Deane, 2013; Hearst, 2000), and one overarching critique is that automated approaches do not capture the complete writing construct. Computers can only respond to features of writing that can be automatically detected, which might exclude nuanced and subjective dimensions that even humans find difficult to assess (Deane, 2013). For instance, the National Council of Teachers of English released a position statement in 2013 (Anson et al., 2013) arguing that "computers are unable to recognize or judge those elements that we most associate with good writing (logic, clarity, accuracy, ideas relevant to a specific topic, innovative style, effective appeals to audience, different forms of organization, types of persuasion, quality of evidence, humor or irony, and effective uses of repetition, to name just a few)."

Negative perceptions of AWE may have consequences with respect to adoption, use, and effectiveness. Users make decisions about technology adoption based on ease of use and utility, and software that is difficult to access, incomprehensible, or seems to provide few benefits may be rejected (Vinkatesh & Davis, 2000). Ertmer and colleagues (Ertmer, 1999; Ertmer, Ottenbreit-Leftwich, Sadik, Sendurur, & Sendurur, 2012) have specifically studied barriers to teachers' use of educational technologies. Logistical barriers, such as a poor student-to-computer ratio or a lack of reliable Internet access, may be substantial when hundreds of students are simultaneously writing and submitting essays to web-based AWE services. Other barriers are grounded in teachers' beliefs, such as the belief that certain domains (e.g., writing) cannot be taught using automated approaches. Instructors who possess necessary resources and expertise (Koehler & Mishra, 2009; Voogt, Fisser, Pareja Roblin, Tondeur, & van Braak, 2013) may nonetheless reject AWE due to skepticism (Curran, Draus, Maruschock, & Maier, 2014).

Students also express mixed views about AWE (Grimes & Warschauer, 2010; Warschauer & Grimes, 2008). In one study (Grimes & Warschauer, 2010), students rated *My Access* favorably in terms of ease, enjoyment, usefulness, and fairness, and reported that they revised more and increased their confidence after using the system. However, students also focused their attention on low-level writing feedback and were sometimes overwhelmed by the amount of feedback. McNamara and colleagues (Roscoe & McNamara, 2013; Roscoe, Allen, Weston, Crossley, & McNamara, 2014) conducted similar feasibility assessments of early versions of *Writing Pal* (W-Pal), an intelligent tutoring system for writing instruction. Their study was conducted in several high school English classrooms over one school year. After students had interacted with W-Pal for several months, the researchers probed their perceptions of the feedback system. Most students (about 80%) rated the writing tools as easy to use, but some students critiqued the system with regards to quantity of feedback (i.e., either too much or not enough). About 60% of students found the feedback to be easy to understand, and about 40% reported that the feedback was "often" or "always" useful. In open-ended responses, students noted that a lack of feedback specificity and personalization were key concerns.

Overall, students and teachers seem to cautiously embrace the potential of AWE for providing summative and formative feedback on writing, but also express doubts regarding scoring accuracy, specificity and personalization, clarity, and quantity of the feedback. When students' preferences for human versus automated feedback have been probed directly, students tended to prefer comments from teachers or peers rather than computers (Curran, Draus, Maruschock, & Maier, 2013; Lai, 2010; Lipnevich & Smith, 2009).

The current study explores college students' perceptions of automated essay scoring and feedback, and examines the effects of perceptions on writing performance, writing process (i.e., revising), and future intentions toward the technology. Specifically, we assess students' initial expectations about AWE scoring and feedback, immediate reactions to received scores and feedback, and final impressions. Importantly, we also manipulate how system capabilities are presented to student users. As noted above, teachers possess conflicting views about the validity of AWE and may communicate these views to their students (Li, Link, & Hegelheimer, 2015). With this manipulation, we can consider how messaging from authority figures—in this case, developers and researchers associated with the AWE itself—might influence user perceptions and outcomes. However, the actual functioning or quality of the system is not manipulated; all students interact with the same scoring and feedback tools.

1.1. Research questions

Research Question 1 (RQ1). When AWE scoring and/or feedback capabilities are presented as either "well established" versus a "work in progress," how does this presentation influence students' expectations about software performance, immediate perceptions of feedback received on their own work, and final impressions of the system?

Research Question 2 (RQ2). Do differences in presentation, expectations, and experience contribute to positive or negative shifts in final perceptions of the system? One possibility is that initial expectations strongly anchor subsequent interpretations of the system. Alternatively, direct experiences and interactions might override original expectations.

Research Question 3 (RQ3). How do positive and negative perceptions of AWE influence writing behaviors and future intentions regarding the system? In terms of writing behaviors, we examine how perceptions of the software and feedback quality relate to revising. Students who believe that automated feedback is more accurate, relevant, or useful may be more inclined to use that feedback and revise extensively. With future intentions, we test whether perceptions influence willingness to use the system again or recommend it to a friend.

2. Method

2.1. Participants

We recruited 110 undergraduate students enrolled in Introduction to Psychology courses at a large university in the southwestern United States. Students received course credit for their participation. Demographically, 35.5% of students self-identified as female with an average age of 22 ($M = 21.8$, $SD = 5.7$). Most students self-identified as Caucasian (40.0%) or Hispanic (20.9%) although other races and ethnicities were represented (see Table 1). Most students spoke primarily English (69.1%) or were fluent in English and another language (28.2%). A small number of students reported another language as their primary language (2.7%) but possessed sufficient English proficiency to participate. Academically, students reported a relatively high average GPA of about 3.5

Table 1
Summary of demographic and academic background across students.

	% of Students	Mean	*SD*
Demographic Information			
Age		21.8	5.7
Gender			
Female	35.5		
Male	64.5		
Race/Ethnicity			
African	1.8		
African-American	6.4		
Asian	2.7		
Caucasian	40.0		
Hispanic	20.9		
Middle Eastern	7.3		
Native American	2.7		
Pacific Islander	2.7		
Multi-ethnic	11.8		
Not Reported	3.6		
Primary Language(s)			
English	69.1		
Multilingual	28.2		
Non-English	2.7		
Academic Information			
Grade-point Average		3.50	0.46
Academic Major Category			
Aviation	20.9		
Biomedical	16.4		
Business	23.6		
Computing	19.1		
Engineering	8.2		
Social Science/Humanities	10.0		
Undecided	1.8		

(*SD* = 0.46) and were enrolled in diverse academic majors. The most common majors were business-related (23.6%), aviation-related (20.9%), computing-related (19.1%), and biomedical-related (16.4%). Table 1 provides further details.

Students self-reported general attitudes toward writing and computing (see Table 2) by rating their agreement to a variety of statements (e.g., "Writing skills are important for success") on a scale from 1 (strongly disagree) to 6 (strongly agree). Overall, students reported positive beliefs about their own writing and computing skills, and endorsed the idea that writing skills are important. These data suggest that students neither hated nor loved writing but did understand its value, and they were comfortable with writing on computers. Most students had no prior experience with automated grading (82.7%) or writing instruction (84.5%) technologies.

Finally, it is worth noting that although this study sampled college writers, W-Pal was designed for high school adolescent writers (see Section 2.2). However, there are several reasons why studying college students is both valid and informative. First, many undergraduates are not necessarily skilled writers (Kellogg & Raulerson, 2007; MacArthur, Philippakos, & Ianetta, 2015). College students may be more sophisticated or proficient than high school adolescents (Crossley, Weston, Sullivan, & McNamara, 2011), but not always by a large margin. Thus, W-Pal's scoring system and feedback should still be appropriate for this population. Second, we were interested in whether college undergraduates could effectively use a system such as W-Pal. It is valuable for educational technologies to test their "reach" and, in terms of sustainability, educational technologies should strive to serve as many learners as possible. If W-Pal is useful for college students, this might be a valuable finding given that college level writing expectations may be more demanding than at the high school level.

Table 2
Summary of prior Attitudes and experience with writing and computing.

	% of Students	Mean	*SD*
Attitudes			
Importance of Writing Skills		5.2	0.7
Enjoyment of Writing		3.5	1.3
Possess Good Writing Skills		4.2	0.9
Like to Write with Pen and Paper		4.0	1.5
Prefer to Writing on Computer		4.4	1.4
Skilled at Using Computers		4.7	1.1
Often have Writing Assignments		4.7	1.1
Writing Technology Experience			
No Grading Software Experience	82.7		
No instructional Software Experience	84.5		

2.2. The writing task and Writing Pal (W-Pal)

Students wrote and revised an essay on the topic of "STEM and psychology in the media" using W-Pal. Students were instructed to answer a target question and "try to convince others to agree with your point of view" and to support their ideas "with logical arguments, examples, and evidence." Students were allotted 20 min to compose original drafts and submit them for automated scoring and feedback, and then given another 10 min to revise after reviewing their feedback. The specific writing prompt appears below:

> Many television shows and movies use stories and ideas drawn from science, technology, engineering, and math. For example, characters might use their knowledge of medicine and psychology to investigate crime scenes, or may use their knowledge of physics and electronics to build or repair important devices. However, some experts have argued that these portrayals of science, technology, engineering, or math are often factually wrong. Think about the concepts and principles that you are learning in your psychology course. *Do television shows and movies accurately portray psychology?*

W-Pal is an intelligent tutoring system that supports writing instruction via strategy instruction, game-based strategy practice, and essay writing practice with automated formative feedback (Roscoe & McNamara, 2013; Roscoe et al., 2014). The primary audience for this software is high school adolescents engaged in argument-based writing. The researchers consulted with high school teachers and student users during development, and studies have primarily focused on high school student populations (Roscoe et al., 2014). Students at the high school level are expected to develop skill with communicating claims, use of supporting logic and evidence, crafting cohesive and coherent texts, and establishing a formal and objective tone (Common Core State Standards, National Governors Association, 2010). Although persuasive writing is only one of several crucial genres that students must master, it holds a prominent place in high school curricula. Many of the core processes and concepts also apply across genres and undergird the advanced writing that students are asked to produce in college.

To teach students about writing strategies, W-Pal includes multiple modules that span prewriting, drafting, and revising aspects of writing (see Hayes, 2012). Each module comprises four to five lesson videos in which animated characters explain and demonstrate writing strategies. For example, the *Body Building* module emphasizes the "CASE" mnemonic for body paragraphs. Students are taught how to present a Concise Argument (i.e., topic sentence) along with Supporting Evidence. Completing one or more lesson videos unlocks educational mini-games that allow

students to practice targeted strategies.

In this study, students only interacted with the automated scoring and feedback tools. W-Pal allows students to practice writing prompt-based, argumentative essays on a variety of topics. A word processing interface allows basic text formatting along with the ability to review the prompt and take notes in a "scratch pad" (see Fig. 1). Essays are also expected to begin with a relevant introduction, include supporting paragraphs, and provide a clear conclusion.

Student essays submitted to W-Pal receive a holistic score on a 6-point scale from 1 ("Poor") to 6 ("Great") similar to the SAT scoring rubric. Each essay also receives automated formative feedback that provides recommendations for proficient writing and strategies corresponding to the lessons. This feedback includes information about writing concepts and procedures along with actionable steps for improvement. For example, short essays that are judged to be lacking in elaboration may receive feedback on how to generate ideas. Similarly, essays that exhibit repetitiveness might receive feedback on improving word choice via paraphrasing. This kind of formative feedback is essential to students' writing development because it communicates the knowledge and methods necessary for growth (McGarrell & Verbeem, 2007; Shute, 2008; Sommers, 1982).

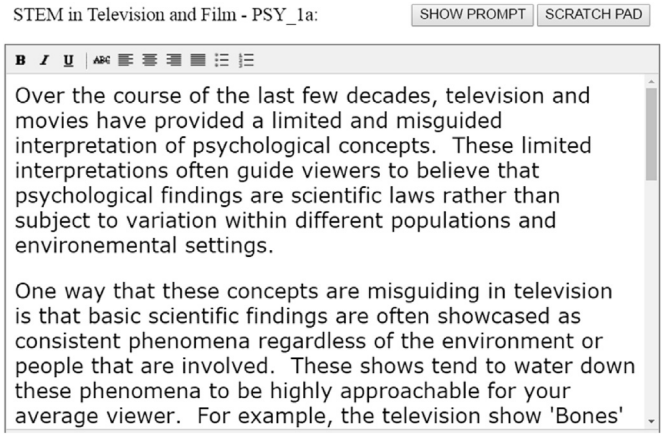

Fig. 1. Screenshot of essay authoring interface.

In W-Pal, categories of formative feedback are implemented via a sequential, hierarchical series of algorithms assessing text *legitimacy*, *length*, *structure*, *introduction* quality, *body* quality, and *conclusion* quality. At the highest level (i.e., if all other categories meet criteria) students receive general *revising* tips. Importantly, feedback is given at the category level rather than specific words, sentences, ideas, or arguments. Each category is associated with a pool of relevant messages that is selected at random (without replacement) if that category is flagged as a problem. Thus, a student whose essay contains weaker body paragraphs might receive strategy suggestions for strengthening their evidence, but these messages are not directly linked to the unique content of the essay. W-Pal always offers one feedback message on an initial topic (i.e., the first problem detected in the sequential hierarchy), and students can choose to receive more feedback on that topic and/or the second problem detected in the hierarchy. Fig. 2 provides an example feedback report from an original essay draft that earned a "Weak" rating (i.e., a score of "2") and received feedback on body-building and introduction-building strategies. After revising, the student earned an "Okay" rating (i.e., a "4").

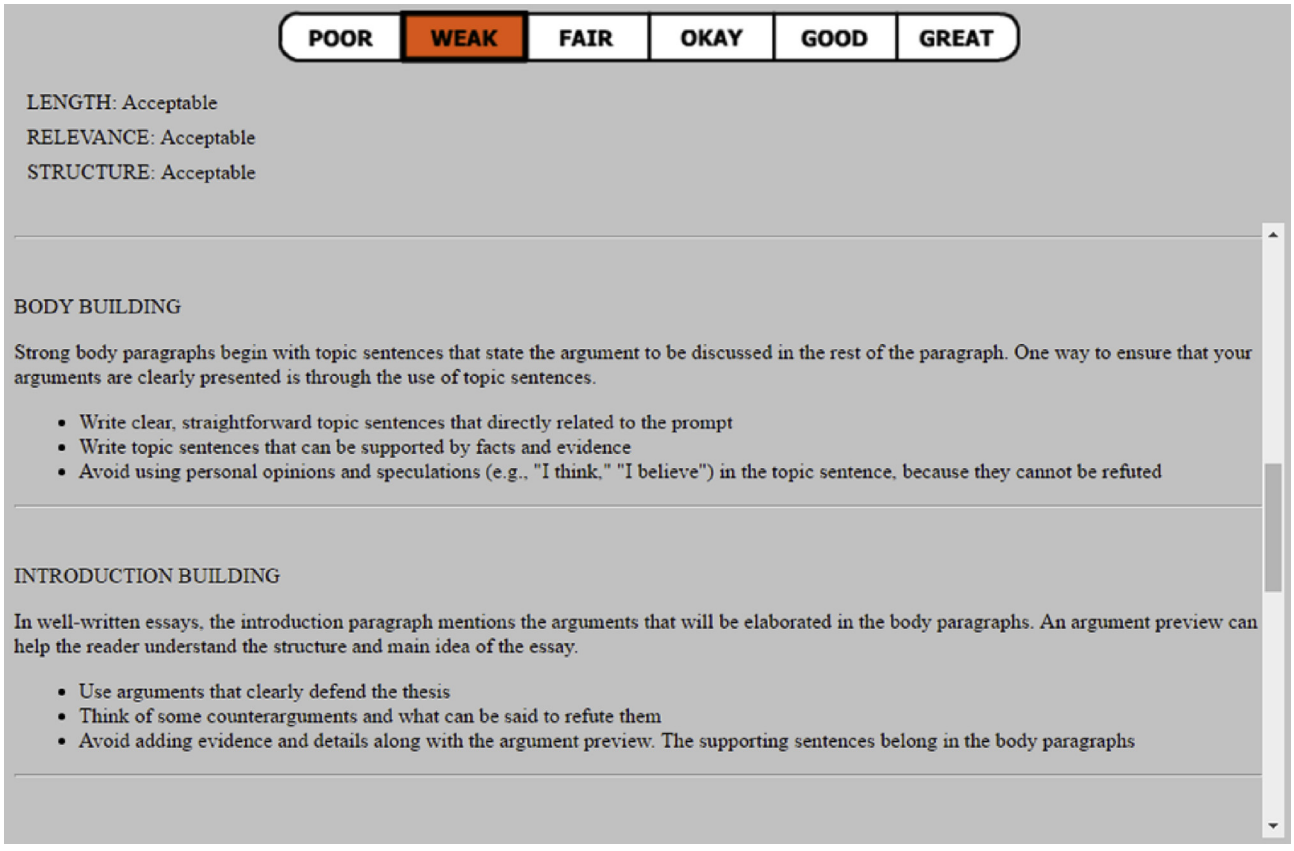

Fig. 2. Screenshot of a W-Pal feedback report including the summative rating and formative feedback messages.

These holistic scoring and feedback processes are driven by a series of NLP algorithms that evaluate diverse lexical, syntactic, semantic, and rhetorical properties of the text, which are driven by a combination of analytical methods and tools (Crossley, Kyle, & McNamara, 2016; Kyle & Crossley, 2015; McNamara, Crossley, & Roscoe, 2013; McNamara, Crossley, Roscoe, Allen, & Dai, 2015; McNamara, Graesser, McCarthy, & Cai, 2014). Algorithms incorporate basic text properties, such as numbers of words and paragraphs, along with more advanced measures. For instance, lexical sophistication is assessed with respect to features such as word diversity, concreteness, frequency, and specificity (Fellbaum, 1998; Kyle & Crossley, 2015). Parts of speech tagging identifies the incidence of nouns, pronouns, and verbs, and different types of each. Similarly, semantically themed word lists assess the occurrence of academic language and concepts related to thoughts, feelings, actions, and beliefs (Tausczik & Pennebaker, 2010). At a higher level, indices assess local, global, and text cohesion (Crossley et al., 2016), such as word stem overlap and semantic similarity across paragraphs (Foltz, 2007), and a narrativity index evaluates the extent to which story-like elements are present in the text (McNamara et al., 2014).

Collectively, linear regression and discriminant function analyses combine indices in algorithms that generate holistic scores along with indicators for feedback categories. Past evaluations have found W-Pal's scoring algorithms to demonstrate acceptable accuracy (e.g., McNamara et al., 2013, 2015). That is, computer-assigned scores are often an exact match or within one point of human-assigned scores. In addition, prior studies have found that high school student users of W-Pal demonstrate modest gains in essay quality and strategy knowledge, and engage in substantive revising (Allen et al., 2014; Crossley, Roscoe, & McNamara, 2013; Roscoe, Jacovina, Harry, Russell, & McNamara, 2015; Roscoe, Snow, Allen, & McNamara, 2015). Feasibility assessments conducted throughout W-Pal development have documented both positive perceptions and critiques that have guided software design (Roscoe et al., 2014).

2.3. Conditions

Students were randomly assigned to one of four experimental conditions that manipulated the presentation of system capabilities. Specifically, students were told that the *automated scoring capabilities* of the system were either well-established or under development. Similarly, students were told that the *automated feedback capabilities* of the system were either well-established or under development. We communicated "strong" features by emphasizing that the development team included multidisciplinary researchers who collaborated for over five years to refine the software, and that research studies had observed good scoring and efficacy (e.g., improved essay quality). We communicated "weak" features by emphasizing the difficult endeavor of AWE, and explained that we were identifying and fixing system "bugs" to improve functionality. The Appendix provides the scripts that were verbally delivered to students.

Importantly, this study did not use deception. It is true that the W-Pal software was developed by a multidisciplinary team, has been evaluated in several studies, and has shown benefits for strategy acquisition, writing performance, and writing motivation. Nonetheless, it is also true that the software is a "living system"—new data are used to continuously refine scoring and feedback algorithms, feedback delivery, and user experience. Thus, all conditions were presented with accurate descriptions of the software, but those descriptions emphasized either current accomplishments or ongoing challenges.

Using a 2 (presented scoring quality) x 2 (presented feedback quality) design, students were assigned to one of four scenarios: *Strong Scoring/Strong Feedback*, *Strong Scoring/Weak Feedback*, *Weak Scoring/Strong Feedback*, or *Weak Scoring/Weak Feedback*. Differences in demographic, academic, and attitudinal background across conditions were assessed via chi-square or analyses of variance. No significant differences were observed.

2.4. Assessing perceptions of automated feedback

2.4.1. Initial feedback expectations

Immediately after the scripted introduction to W-Pal, students reported background information and their initial expectations about system capabilities using a paper-and-pencil survey. Students responded to four items using a 1 (lowest) to 6 (highest) scale. One item assessed students' *expected scoring accuracy* of the software ("How accurate do you expect the computerized scoring will be in this study?"). Because automated scores were summative in nature, we only asked students to consider accuracy. However, feedback provided by W-Pal is formative, and thus students were asked to consider a broader array of dimensions. Three items assessed *expected feedback accuracy*, *expected feedback relevance*, and *expected feedback usefulness* (e.g., "How relevant do you expect the computerized feedback will be in this study?").

Prior research on feedback perceptions has occasionally observed that multiple feedback ratings load onto a single factor (Strijbos, Narciss, & Dünnebier, 2010). To test this possibility, a factor analysis using principal axis factoring with oblique rotation was conducted with the three feedback rating items (accuracy, relevance, and usefulness). All three items were inter-correlated ($rs > .60$, $p < .001$). A single factor with an eigenvalue greater than 1.00 (eigenvalue = 2.27) emerged that accounted for 75.7% of the variance, and all three items exhibited very high loadings (>.75) on this factor. Thus, for purposes of analysis, we created a single composite score named *expected feedback quality* by averaging the three individual items.

2.4.2. Immediate feedback perceptions

W-Pal embeds an online survey to elicit users' ratings of the scores and feedback they receive. Specifically, W-Pal users rate perceptions *after* they have written an essay, received scores and feedback on their original draft, and attempted to revise, but *before* they receive scores or feedback on their revisions. Thus, immediate perception ratings measure students' reactions to the system in real time without being influenced by how their scores changed. Ratings are made using a 1 (lowest) to 4 (highest) scale. One item assesses perceptions of *immediate scoring accuracy* ("How accurately did Writing Pal score your original essay?"). Three other items assess perceptions of *immediate feedback understandability* ("How understandable was the feedback you received from Writing Pal?"), *immediate feedback relevance* ("How relevant was the Writing Pal feedback for revising your essay?"), and *immediate feedback usefulness* ("How easy was it to use the Writing Pal feedback to revise your essay?").

For the three feedback perceptual ratings, we again tested whether individual items might load on a single factor. Immediate feedback perceptions were positively inter-correlated ($rs > .50$, $p < .001$). A single factor emerged with an eigenvalue greater than 1.00 (eigenvalue = 2.18) that accounted for 72.8% of the variance, and all three items exhibited high loadings (> .70) on this factor. Thus, for purposes of analysis, we created a single composite score named *immediate feedback quality* by averaging the three individual items.

2.4.3. Final feedback perceptions

After finishing assigned tasks in W-Pal, students completed a

follow-up paper-and-pencil survey regarding their perceptions. Because we were specifically interested in students' updated impressions of the software, these four items were carefully worded in relation to *perceptual changes*. One item pertained to changes in perceptions of *scoring accuracy* ("Was the software scoring more or less accurate than you expected?"). Three items pertained to changes in perceptions of *feedback accuracy*, *feedback relevance*, and *feedback usefulness* (e.g., "Was the software feedback more or less relevant than you expected?"). Ratings used a 5-point scale from −2 (much less than expected) to 0 (just as expected) to +2 (much more than expected). These ratings describe the extent to which students' perceptions shifted in a positive or negative direction after interacting with the software.

As above, we considered the possibility that the three feedback ratings behaved as separate judgments or loaded on a single factor. All three feedback perceptual change scores were highly intercorrelated ($rs > .63$, $p < .001$) and loaded on to a single factor (eigenvalue = 2.28) that accounted for 76.1% of the variance. Factor loadings were all greater than .78, and a single averaged composite score was created and named *perceptual change in feedback quality*.

2.4.4. Future intentions

The follow-up survey concluded with two yes-or-no items regarding students' future intentions related to W-Pal. Students were asked, "Would you use this software again to help you improve your writing?" and "Would you recommend this software to a friend who needed writing help?" These items broadly capture students' willingness to continue using or interacting with the software after the conclusion of the study.

2.5. Assessing writing quality and revising behavior

2.5.1. Essay quality

The quality of students' original and revised drafts was assessed via scoring algorithms employed by W-Pal. W-Pal scores have been previously shown to demonstrate acceptable agreement with human raters (McNamara et al., 2014, 2015). Essay scores are reported using a 6-point scale ranging from 1 (lowest quality) to 6 (highest quality).

2.5.2. Revising

To assess revising, we adapted coding procedures reported in previous research (Bridwell, 1980; Crawford, Lloyd, & Knoth, 2008; Faigley & Witte, 1981). Researchers prepared the essays by collaboratively segmenting texts to identify each edit. Original and revised drafts were inspected using the document comparison function of a word processing program, and then each change from original to revised draft was tagged. These revisions could occur at three levels. Mechanics-level revisions comprised edits related to conventions, punctuation, spelling, and grammar (e.g., reconjugating a verb from past tense to present tense). Word-level revisions comprised edits within single words or brief phrases, such as replacing a vague term ("furniture") with a more precise term ("armoire"). Finally, sentence-level revisions occurred when entire sentences or clauses were edited. For example, writers might revise their thesis statement to better preview key arguments, or might remove a colloquial comment from the essay. After tagging each edit, we subsequently coded for a) specific actions taken to revise the text and b) whether these edits preserved or altered the meaning of the surrounding text.

Writers can edit their texts by adding information, deleting or removing details, substituting or replacing content, or reorganizing portions of the text. *Additions* are revisions in which new text is inserted, such as new evidence to support an argument. *Deletions* represent the opposite case when text is removed without replacement. *Substitutions* occur when text is not merely added or removed but is replaced with alternate text. In the example above, replacing the word "furniture" with "armoire" would be a substitution. Finally, *reorganizations* represent the redistribution, consolidation, or other movement of text from one section of the essay to another. For instance, an author may decide that an example might be more memorable if presented at the end of the essay and thus move that example to the conclusion.

To assess the reliability of coding revision actions, a random sample of 30 essays was extracted that contained 407 distinct revisions. Two researchers independently categorized the action taken in each revision. Reliability of coding was high ($\kappa = .92$). After discrepancies were resolved, a single researcher completed the remaining coding.

Finally, revisions can either preserve or alter the meaning of the surrounding text. *Superficial edits* maintain the current meaning of the text. For instance, many changes to punctuation or spelling do not affect essay content. Likewise, writers might reorganize a few words or sentences to improve readability without changing the ideas communicated. In contrast, *substantive edits* change the claims and concepts expressed in the text. Authors might add or substitute information that changes the interpretation of events (e.g., "the woman *played* in the basketball game" versus "the woman *competed* in the *championship* basketball game"). Writers could also realize that a piece of evidence is ambiguous and remove it from the essay.

To assess the reliability of coding impact, another random sample of 30 essays was extracted that contained 398 distinct revisions. Two researchers independently categorized the impact of each revision on the meaning of surrounding text. Reliability of coding was acceptable ($\kappa = .81$). After discrepancies were resolved, a single researcher completed the coding.

3. Analysis and results

3.1. Overall perceptions of automated scoring and feedback in Writing Pal

Analyses first examined students' expectations, immediate perceptions, and perceptual changes collapsed across conditions. Initial expectation ratings of scoring accuracy were 4.3 ($SD = 0.9$), corresponding to the belief that scoring would be somewhat accurate. Initial expectations of feedback quality were 4.2 ($SD = 0.8$), corresponding to beliefs that the feedback would be somewhat accurate, relevant, and useful. Immediate perceptions of scoring accuracy ($M = 3.2$, $SD = 0.7$) and feedback ($M = 3.2$, $SD = 0.7$) were likewise positive. Finally, students expressed a marginally significant positive shift in their perceptions of scoring accuracy after using W-Pal ($M = +0.2$, $SD = 1.0$), $t(109) = 1.86$, $p = .065$, $d = .16$. Students also expressed a positive shift in their perceptions of feedback quality ($M = +0.3$, $SD = 0.9$), $t(109) = 2.92$, $p = .004$, $d = .24$.

As expected, students' expectations and impressions of W-Pal might best be described as cautiously positive. Such perceptions are typical of findings in the literature (Grimes & Warschauer, 2010) and prior research on W-Pal (Roscoe et al., 2014).

3.2. Effects of presentation on expectations and immediate perceptions

We next examined the effects of presentation on students' initial expectations and immediate perceptions of the system (RQ1). Analyses were conducted as 2 (presented scoring quality) x 2 (presented feedback quality) between-subjects ANOVAs, which allowed

us to consider main effects and interactions among the manipulations. However, no interactions were found between presented scoring and feedback quality manipulations, and thus only main effects are discussed. Table 3 reports means and standard deviations by condition. For completeness, Table 3 also includes means for perceptual change variables, but perceptual change effects are addressed more specifically in Section 3.3.

3.2.1. Effects on initial expectations

Presented scoring quality seemed to impact initial expectations of scoring accuracy. When we emphasized that the careful testing and accuracy of the scoring system, students indeed expected the scoring system to perform more accurately, $F(1, 106) = 7.86$, $p = .006$, partial $\eta^2 = .069$). However, presented scoring accuracy appeared to have no impact on initial expectations about the quality of the automated feedback, $F(1, 106) < 1.00$, $p = .484$, partial $\eta^2 = .005$.

Presented feedback quality had little impact on expectations about scoring accuracy, $F(1, 106) < 2.32$, $p = .131$, partial $\eta^2 = .021$; but exhibited a marginally significant positive trend related to expectations about feedback quality, $F(1, 106) = 3.07$, $p = .084$, partial $\eta^2 = .028$). When we communicated that the feedback system had been carefully designed to be "accurate, on topic, and helpful," and that research had shown that the feedback "can help students raise their score," expectations were slightly more positive.

3.2.2. Effects on immediate perceptions

Presented scoring quality appeared to have no impact on immediate perceptions of scoring accuracy, $F(1,106) = 1.42$, $p = .236$, partial $\eta^2 = .013$; or feedback quality, $F(1, 106) < 1.00$, $p = .875$, partial $\eta^2 = .000$. Presented feedback quality had no impact on immediate perceptions of scoring accuracy, $F(1, 106) < 1.00$, $p = .433$, partial $\eta^2 = .006$), but was marginally related to immediate perceptions of feedback quality, $F(1, 106) = 3.84$, $p = .053$, partial $\eta^2 = .035$. The way in which W-Pal feedback capabilities were introduced may have had a mild influence on how students reacted to the feedback they received.

3.2.3. Summary

Overall, presenting the system in terms of stronger or weaker scoring capabilities, or stronger or weaker feedback capabilities, had a small but noticeable influence. Presented scoring quality only influenced initial expectations about scoring but then appeared to fall away as students interacted with the system. In contrast, presented feedback quality had a small influence on perceptions before and while using the system.

3.3. Effects of presented quality, expectations, and experience on perceptual change

The preceding analyses established that the manner of presentation could affect subsequent expectations and immediate perceptions. We further hypothesized that final perceptions of AWE should also be influenced by expectations and immediate experience within the software (RQ2). We conducted two multiple regressions to predict perceptual changes related to scoring accuracy and feedback quality. Predictors were entered in three blocks to assess the relative value (i.e., R^2 change) for adding each set of variables: a) presented scoring and feedback quality, b) initial expectations of scoring accuracy and feedback quality, and c) immediate perceptions of scoring accuracy and feedback quality (see Table 3 for means).

3.3.1. Predicting perceptual changes of scoring accuracy

Table 4 summarizes the predictive value of presentation, expectations, and immediate perceptions on changes in students' perceptions of scoring accuracy.

Presented scoring accuracy and feedback quality alone (Model 1) were not predictive of students' perceptual changes for scoring accuracy ($R^2 = .01$). Likewise, the combined effects of presentation and initial expectations (Model 2) were also not predictive of perceptual changes ($R^2 = .04$). However, the inclusion of students' immediate reactions to the feedback received on their essays (Model 3) resulted in a significant increase in the variance explained ($\Delta R^2 = .33$), and Model 3 accounted for 37% of the variance in perceptual change. Several variables were significant predictors, including expectations about scoring accuracy ($\beta = 0.24$) and feedback quality ($\beta = -0.38$), and immediate perceptions of scoring accuracy ($\beta = .47$) and feedback quality ($\beta = .26$). Thus, both expectations and direct experience were related to how perceptions of scoring accuracy changed over time.

3.3.2. Predicting perceptual changes of feedback quality

Table 5 summarizes the predictive value of presentation, expectations, and immediate perceptions on changes in students' perceptions of feedback quality.

Presented scoring accuracy and feedback quality alone (Model 1) accounted for a small amount of variance ($R^2 = .05$), although only feedback presentation was a significant predictor ($\beta = .21$). When expectations were included (Model 2), the increase in variance explained was negligible ($\Delta R^2 = .03$). In Model 3, the inclusion of immediate perceptions resulted in a significant increase in the variance explained ($\Delta R^2 = .35$) and Model 3 accounted for 44% of the variance in perceptual change. Two variables were significant

Table 3
Mean ratings of expectations, immediate perceptions, and final perceptions by condition.

Feedback Ratings	Presented Scoring and Feedback Quality Conditions			
	Strong Scoring, Strong Feedback ($n = 26$)	Strong Scoring, Weak Feedback ($n = 27$)	Weak Scoring, Strong Feedback ($n = 28$)	Weak Scoring, Weak Feedback ($n = 29$)
Initial Expectations				
Scoring Accuracy	4.3 (1.1)	4.7 (0.6)	4.0 (0.8)	4.1 (0.8)
Feedback Quality	4.4 (1.1)	4.2 (0.7)	4.3 (0.7)	4.0 (0.7)
Immediate Perceptions				
Scoring Accuracy	3.1 (0.6)	3.2 (0.7)	3.5 (0.6)	3.2 (0.9)
Feedback Quality	3.3 (0.5)	3.1 (0.8)	3.4 (0.6)	3.1 (0.8)
Perceptual Change				
Scoring Accuracy	+0.0 (1.1)	+0.1 (0.9)	+0.4 (1.0)	+0.2 (1.1)
Feedback Quality	+0.4 (0.8)	0.0 (0.9)	+0.5 (0.8)	+0.2 (1.1)

Note. [a]$p \leq .001$. [b]$p \leq .01$. [c]$p \leq .05$. [d]$p \leq .10$. All Feedback Quality ratings are composites computed by averaging individual feedback ratings (see Method). Perceptual Change ratings were given on a scale of −2 to +2 and are not difference scores.

Table 4
Multiple regression predicting perceptual change for scoring Accuracy.

Predictor	Coefficients			Model Fit			Model Change	
	β	t	p	R^2	F	p	ΔR^2	p
Model 1				.01	$F(2, 107) < 1.00$	.567	.01	.567
Presentation								
Scoring Accuracy	−0.10	−1.05	.298					
Feedback Quality	0.02	0.22	.827					
Model 2				.04	$F(2, 105) < 1.00$	.436	.03	.268
Presentation								
Scoring Accuracy	−0.14	−1.40	.166					
Feedback Quality	0.08	0.74	.458					
Initial Expectations								
Scoring Accuracy	0.19	1.54	.128					
Feedback Quality	−0.17	−1.37	.175					
Model 3				.37	$F(2, 103) = 9.99$	< .001	.33	<. 001
Presentation								
Scoring Accuracy	−0.08	−0.98	.332					
Feedback Quality	0.03	0.37	.710					
Initial Expectations								
Scoring Accuracy	0.24	2.29	.024					
Feedback Quality	−0.38	−3.60	< .001					
Immediate Perceptions								
Scoring Accuracy	0.47	5.54	< .001					
Feedback Quality	0.26	2.96	.004					

Note. Presentation variables were entered via dichotomous coding (weak = 0 and strong = 1). All Feedback Quality ratings are composites computed by averaging individual feedback ratings (see Method).

and positive predictors: immediate perceptions of scoring accuracy ($\beta = .18$) and immediate perceptions of feedback quality ($\beta = 56$). Thus, students' final perceptions of feedback accuracy appeared to be determined primarily by their direct experiences with and reactions to the scoring and feedback report from W-Pal.

3.4. Writing, revising, and future intentions: potential consequences of perceptions

Final analyses considered the possible consequences of positive and negative perceptions of AWE scoring and feedback (RQ3). Users who are highly skeptical or critical of automated writing instruction may ignore the feedback received and thus revise less (or less well). Similarly, dissatisfied users may choose to avoid the software in the future. To address these questions, we first examined students' writing quality and revising behaviors, and then conducted correlations between perceptions, essay scores, and revising patterns. We then considered how system perceptions differed between students who expressed willingness or unwillingness to use W-Pal in the future or to recommend it to friends.

3.4.1. Essay quality and revising

Students spent about 18–19 min ($M = 18.7$, $SD = 2.3$) composing their original drafts, which were generally weak to fair in quality ($M = 2.7$, $SD = 0.9$). Students spent about 8–9 min ($M = 8.6$, $SD = 2.1$) revising, and final drafts were generally fair in quality

Table 5
Multiple regression predicting perceptual change for feedback quality.

Predictor	Coefficients			Model Fit			Model Change	
	β	t	p	R^2	F	p	ΔR^2	p
Model 1				.05	$F(2, 107) = 3.03$	.053	.05	.053
Presentation								
Scoring Accuracy	-.10	−1.09	.279					
Feedback Quality	.21	2.21	.029					
Model 2				.08	$F(2, 105) = 2.33$	.060	.03	.205
Presentation								
Scoring Accuracy	−0.08	-.087	.384					
Feedback Quality	0.16	1.56	.121					
Initial Expectations								
Scoring Accuracy	−0.12	−0.98	.328					
Feedback Quality	0.21	1.79	.076					
Model 3				.44	$F(2, 103) = 13.26$	< .001	.35	< .001
Presentation								
Scoring Accuracy	−0.07	−0.88	.383					
Feedback Quality	0.10	1.22	.224					
Initial Expectations								
Scoring Accuracy	−0.01	−0.07	.947					
Feedback Quality	−0.06	−0.57	.570					
Immediate Perceptions								
Scoring Accuracy	0.18	2.25	.027					
Feedback Quality	0.56	6.69	< .001					

Note. Presentation variables were entered via dichotomous coding (weak = 0 and strong = 1). All Feedback Quality ratings are composites computed by averaging individual feedback ratings (see Method).

($M = 3.1$, $SD = 0.9$). Gains in essay quality across drafts were significant, $t(109) = 5.08$, $p < .001$, and showed a moderate level of improvement ($d = .43$).

Students demonstrated substantial variability in their revising behaviors. A few students engaged in almost no revising—the minimum number of revisions for all categories was zero. However, most students engaged in at least some degree of revising and some students revised extensively (Table 6). Within each revision category (i.e., level, action, and impact), certain types of revisions were more common than others. A series of pair-wise t-tests within each category (adjusted alpha = .013) were conducted to assess relative mean frequencies. Overall, students were similarly likely to implement word-level and sentence-level revisions ($p = .015$), but mechanics-level revisions occurred less often ($ps < .001$). In terms of revising actions, additions occurred most often, followed by substitutions, deletions, and reorganizations (all $ps < .001$). Superficial and substantive revisions occurred with similar frequency ($p = .303$).

Correlations tested the relationships between revisions and changes in essay scores from original to revised drafts. The total number of revisions was positively and significantly related to gains in essay quality ($r = .21$, $p = .031$). Moreover, revisions that added new content ($r = .32$, $p = .001$) or substantively changed the meaning of surrounding text ($r = .29$, $p = .002$) were also associated with improvement. Revisions at the level of mechanics or words, revisions that removed content or restructured content, and revisions that did not change the meaning of the text, appeared to have little systematic relationship with essay gains.

These results suggest that college undergraduates were able to effectively use W-Pal to write and revise their essays. After receiving feedback, students' essays improved moderately in quality. Although this study was not designed to establish the efficacy of W-Pal for college-age populations, these data do speak to the viability and value of future research that explores W-Pal deployment with older students.

3.4.2. Correlations between perceptions, essay quality, and revising behaviors

Correlations were conducted between gains in essay score, all revising behaviors, immediate perceptions, and perceptual changes. Immediate perception variables allow us to consider how writing and revising varied based on students' reaction to the feedback. For instance, did students revise more carefully when they perceived the feedback to be better? Perceptual change ratings allow us to explore whether positive or negative shifts in perceptions were aligned to writing quality and revising. Did students revise less if they were disappointed in the accuracy of scoring (i.e., a negative perceptual change)?

Gains in essay quality were uncorrelated with immediate perceptions of scoring accuracy ($r = -.08$, $p = .40$), immediate perceptions of feedback quality ($r = .08$, $p = .375$), or perceptual change in scoring accuracy ($r = .12$, $p = .22$). Improvement in essay quality was marginally related to perceptual changes in feedback quality ($r = .18$, $p = .059$). Similar trends were observed for correlations between perceptions and original draft scores (all $rs < .11$) and revised draft scores ($rs < .18$), with one exception. Overall, students' ratings of scoring accuracy or feedback quality were not driven by the scores received—students were not merely "rewarding" high scores and "punishing" low scores. However, perceptual changes in feedback quality were correlated with higher revision scores ($r = .20$, $p = .033$). Students who earned a higher score on their final draft were more likely to perceive feedback more favorably than they expected.

No significant correlations were found between immediate perception ratings, perceptual changes, and any revising behavior (all $rs < .17$ in magnitude, and all $ps > .09$). Students seemed to revise no differently based on whether they perceived the feedback as more or less accurate, relevant, or useful, and regardless of whether they perceived the feedback and scoring as better or worse than expected.

Table 6
Mean frequency of revision Actions and correlations with score gains.

	Minimum	Maximum	Mean	*SD*	*r* with essay score gain
Total Revisions	0.0	41.0	11.6	8.1	.21[c]
Revision Level					
Mechanics	0.0	10.0	1.7	2.3	.05
Word	0.0	32.0	4.0	5.6	.06
Sentence	0.0	33.0	5.9	4.9	.25[b]
Revision Action					
Addition	0.0	23.0	6.6	4.3	.32[a]
Deletion	0.0	26.0	1.6	3.0	-.02
Substitution	0.0	24.0	3.0	4.2	.06
Reorganization	0.0	6.0	0.3	0.8	.14
Revision Impact					
Superficial	0.0	36.0	5.4	6.7	.05
Substantive	0.0	33.0	6.2	4.8	.29[b]

Note. [a]$p \leq .001$. [b]$p \leq .01$. [c]$p \leq .05$. [d]$p \leq .10$.

3.4.3. Perceptions and future intentions

Out of all 110 students, 68.2% reported that they would be willing to use W-Pal again. Out of 109 students (i.e., one student did not respond), 70.9% reported that they would recommend W-Pal to a friend. Thus, over two-thirds of students found W-Pal to be valuable or useful enough to express positive future intentions. However, a meaningful number of students expressed a lack of willingness to do so (Table 7).

Strong perceptual differences were observed for students who were willing to use W-Pal again in the future. On average, students who responded "Yes" held more positive initial expectations than students who responded "No." Similarly, students willing to reuse W-Pal rated immediate perceptions of scoring accuracy and feedback quality more favorably, and also reported a positive shift in their perceptions of scoring and feedback. Willingness to use the software again was associated with a positive attitude throughout the study. A very similar pattern was observed for students who were willing to recommend W-Pal to a friend.

The effects of initial expectations, immediate perceptions, and perceptual change on willingness to use W-Pal again were further explored using logistic regression to predict students' "yes" or "no" decisions (Table 8). The overall model demonstrated decent fit, $X^2(8) = 75.27$, $p < .001$, Nagelkerke $R^2 = .69$. Only one predictor was statistically significant. Students who expressed a positive change in their perceptions of feedback quality were more likely to indicate willingness to use W-Pal again in the future. Specifically, given a one unit increase in students' perceptions of feedback quality, and after adjusting for the effect of the other variables in the model, students' probability of agreeing with this item was .90 $\left(P = \frac{e^B}{1+e^B}\right)$.

A similar analysis was conducted to predict students' willingness to recommend the software to friend (Table 9). The model demonstrated decent fit, model $X^2(8) = 80.60$, $p < .001$, Nagelkerke $R^2 = .75$. Two predictors were significant. Students who held more positive initial expectations of feedback quality were somewhat less willing to recommend W-Pal to their friends but the effect was small. Given a one unit increase in students' expectations of feedback quality, and after adjusting for the effect of the other variables in the model, students' probability of agreeing with this item was only .07 $\left(P = \frac{e^B}{1+e^B}\right)$. One possibility is that students who had

Table 7
Mean perceptual rating differences and significance tests related to future intentions for W-Pal use.

Rating	Indicators of Future Intentions									
	Use W-Pal Again?					Recommend to a Friend?				
	No ($n = 35$)	Yes ($n = 75$)	$F(1,108)$	p	η^2_p	No ($n = 31$)	Yes ($n = 78$)	$F(1,107)$	p	η^2_p
Initial Expectations										
Scoring Accuracy	4.0 (1.2)	4.4 (0.7)	4.70	.032	.04	3.7 (1.1)	4.4 (0.8)	6.15	.015	.05
Feedback Quality	3.8 (0.9)	4.4 (0.6)	20.83	< .001	.16	3.7 (1.0)	4.4 (0.6)	22.30	< .001	.17
Immediate Perceptions										
Scoring Accuracy	2.8 (0.9)	3.4 (0.6)	21.08	< .001	.16	2.7 (0.8)	3.4 (0.6)	22.49	< .001	.17
Feedback Quality	2.7 (0.8)	3.5 (0.5)	38.27	< .001	.26	2.6 (0.8)	3.5 (0.5)	48.81	< .001	.31
Perceptual Change										
Scoring Accuracy	−0.4 (0.9)	+0.4 (1.0)	17.20	< .001	.14	−0.4 (1.0)	+0.4 (1.0)	13.20	< .001	.11
Feedback Quality	−0.6 (0.7)	+0.7 (0.8)	67.02	< .001	.38	−0.7 (0.7)	+0.6 (0.7)	72.24	< .001	.40

Note. All Feedback Quality ratings are composites computed by averaging individual feedback ratings. Perceptual Change ratings were given on a scale of −2 to +2 and are not difference scores.

Table 8
Logistic regression predicting willingness to use W-Pal in the future.

Predictor	Coefficients				
	B	*SE*	*Wald*	*p*	e^B
Presentation					
Scoring Accuracy	−0.61	0.40	2.34	.126	0.55
Feedback Quality	0.03	0.36	0.01	.929	1.03
Initial Expectations					
Scoring Accuracy	0.91	0.62	2.16	.141	2.48
Feedback Quality	−1.46	1.02	2.04	.154	0.23
Immediate Perceptions					
Scoring Accuracy	1.03	0.66	2.45	.117	2.79
Feedback Quality	0.61	0.63	0.94	.332	1.83
Perceptual Change					
Scoring Accuracy	0.14	0.54	0.06	.799	1.15
Feedback Quality	2.20	0.63	12.24	< .001	9.03

Note. All Feedback Quality ratings are composites computed by averaging individual ratings (as explained in the Method). Perceptual Change ratings were given on a scale of −2 to +2 and are not difference scores.

Table 9
Logistic regression predicting willingness to recommend W-Pal to a friend.

Predictor	Coefficients				
	B	*SE*	*Wald*	*p*	e^B
Presentation					
Scoring Accuracy	0.11	0.45	0.06	.814	1.11
Feedback Quality	0.15	0.41	0.13	.722	1.16
Initial Expectations					
Scoring Accuracy	1.17	0.67	3.06	.080	3.22
Feedback Quality	−2.58	1.21	4.54	.033	0.08
Immediate Perceptions					
Scoring Accuracy	1.33	0.72	3.34	.065	3.77
Feedback Quality	1.10	0.73	2.26	.133	3.00
Perceptual Change					
Scoring Accuracy	−0.21	0.63	0.11	.739	0.81
Feedback Quality	2.848	0.78	13.32	< .001	17.25

Note. All Feedback Quality ratings are composites computed by averaging individual feedback ratings (see Method). Perceptual Change ratings were given on a scale of −2 to +2 and are not difference scores.

inflated expectations of the system were more likely to be disappointed in its actual performance. The best predictor was again a positive shift in perceptions of feedback quality. Students who rated the feedback as higher quality than expected (i.e., more accurate, relevant, and useful) were more willing to recommend the software to others. Given a one unit increase in students' perceptions of feedback quality, and after adjusting for the effect of the other variables in the model, students' probability of agreeing with this item was .94 $\left(P = \frac{e^B}{1+e^B}\right)$.

4. Discussion

As AWE technologies gain in popularity and use, it is important to continue improving their utility and effectiveness (Stevenson & Phakiti, 2013). The current study examined students' expectations and perceptions of automated scoring and feedback, and explored the effects of these impressions on writing quality, revising behaviors, and future intentions. Students (and teachers) often hold mixed opinions about AWE (Grimes & Warschauer, 2010). Although favorable perceptions of AWE accuracy, relevance, and usefulness might lead students to "make the most" of computer-based support, deep skepticism or dislike could result in ignoring automated feedback or rejecting the system (Vinkatesh & Davis, 2000). One way to improve AWE efficacy may to better understand the formation and impact of students' perceptions, and explore how these impressions might be productively managed.

Students' perceptions of W-Pal scoring and feedback were positive overall. Students seemed open to the idea that a computer could score their writing accurately and provide feedback that might help them improve. The major findings were that presentation, expectations, and direct experience all seemed to influence students' perceptions of AWE. These different sources of system judgments may also have had a cascading effect. That is, the manner in which a system is initially presented may plant the seed for more positive or negative expectations, which in turn color subsequent human-computer interactions and intentions. However, results suggest that direct experiences perhaps mattered most. A productive interaction with AWE tools (e.g., believing that one has received accurate and useful scores and feedback) may override initial doubts. In the following discussion, we consider findings pertaining to individual research questions, along with limitations and directions for future research.

4.1. Effects of presentation of AWE capabilities

We first considered how students' perceptions might be influenced by the manner in which software features were presented (RQ1). This is important because teachers' attitudes toward AWE may directly or indirectly influence students' perceptions and interactions (Li et al., 2015). Educational researchers have previously observed how teachers' beliefs about students and learning can affect their teaching behaviors, and students can internalize these attitudes as self-beliefs or implicit theories (Yeager & Dweck, 2012). Similarly, studies of collaborative learning and classroom discourse have found that teachers' practices and expectations are emulated

by their students (Webb, Nemer, & Ing, 2006). When teachers engage in unelaborated explaining and questioning, students mirror such "knowledge transmission" when helping each other in small groups. Given that teachers have a substantial effect on students' beliefs and behaviors, it is plausible that their attitudes and approaches toward AWE may also carry over.

We found that presentational claims about the accuracy and quality of the formative feedback had a modest influence on students' perceptions throughout the study. When informed that W-Pal's feedback system had been developed by experts to be accurate, on topic, helpful, and understandable, students held somewhat more positive expectations about the system, reacted to feedback received more favorably, and concluded the study with a more favorable impression of W-Pal. In contrast, portraying the feedback system as a work in progress, with errors and "bugs" yet to be found and fixed, was related to less positive perceptions.

These results suggest that system presentation may indeed be a meaningful consideration when introducing and deploying AWE software in classrooms. If educators present AWE as a valid or useful tool, students may be more inclined to receive the system favorably and focus on potential benefits. However, if teachers present the software in a highly critical manner, this may sour students' feelings against the system throughout subsequent writing activities. These negative perceptions could, in turn, cause students to disengage with the software and perhaps receive less benefit. In some cases, lower efficacy of AWE might be a self-fulfilling prophecy that stems from early negative expectations.

Importantly, in this study, system presentation was manipulated by representatives of the "development team" rather than teachers. In principle, one should heed the claims researchers or developers describing the capabilities of their own software. However, teachers have many opportunities and perhaps more authority to communicate their attitudes to students. Thus, it is an empirical question as to whether developers' claims have more or less impact on students' beliefs than a respected instructor. Yet another factor may be "word of mouth" evaluations and recommendations from peers. If a program is popular among students, they may share their enthusiasm and engagement with each other. Negative recommendations or warning from peers, in contrast, might undermine otherwise positive messages from developers or teachers. For these reasons, it is unknown whether presentation effects in this study are stronger or weaker than what would be observed in an authentic classroom setting. In future studies, it may be necessary to test the effects of different presenters by contrasting developers' claims, teachers' framing, and students' word of mouth evaluations. Congruency across presenters could have additive effects, but a more interesting question is what happens when presenter opinions are in conflict.

4.2. Effects of expectations and experience

At the end of the study, students indicated whether the software exceeded their expectations, met their expectations, or was disappointing. We examined how presentation, expectations, and immediate perceptions predicted positive or negative changes these final impressions (RQ2). Results suggest that expectations and experience both influenced perceptual change, with perhaps a greater influence stemming from immediate experiences.

For perceptual changes of scoring accuracy, all expectations and immediate perceptions ratings were significant predictors. Positive initial expectations about scoring accuracy may have set students' up to be more receptive (or forgiving) to the scores they received. Students who believed from the outset that scoring would be accurate may have been more likely to take scores at face value or appreciate the summative information communicated. Interestingly, positive initial expectations about feedback quality were *negatively* related to final perceptions of scoring accuracy. This pattern may represent a "violation of expectations" or tradeoff effect in students' judgments of AWE scoring. Students who expected feedback quality to be very high may have been disappointed or surprised when automated scoring did not seem to align. That is, high expectations of feedback quality may result in a similarly high bar set for scoring accuracy, which may lead students to be more sensitive to perceived inaccuracies.

Positive immediate perceptions of scoring and feedback also contributed to positive shifts in perceptions of scoring accuracy. When students believed that the software communicated appropriate scores and recommendations for their essays, students were more likely to view the scores as more accurate than expected, too. It is interesting that perceptions of feedback carried over to perceptions of scoring. One possibility is that valid feedback lends weight to the perceived validity of the scores. Students might be skeptical that a computer can score essays correctly, but if they receive written feedback that appears to pinpoint authentic strengths and weaknesses in their essays, it may suggest that the system is indeed capable of meaningful assessment. Conversely, feedback that was perceived to be incorrect or off-topic may contribute to further skepticism of the scoring capabilities of an AWE.

With regard to perceptual changes of feedback quality, the best predictors were students' immediate perceptions. After submitting their revised essay to W-Pal, but before seeing any scores for that revised draft, students were asked to rate how accurately their original draft had been scored and the quality of the feedback they had received. Students who felt that W-Pal had graded their writing fairly and given good feedback were more likely to report a positive shift in perceptions (i.e., better than expected). In contrast, receiving inaccurate scores, off-topic feedback, or unusable feedback may have inspired disappointment. This result suggests that students' direct experiences with AWE scoring and feedback have the most meaningful impact on whether they are satisfied or dissatisfied—initial expectations are less important.

One remaining question is why initial expectations influenced perceptual changes for scoring accuracy but not feedback quality. One possibility is that students' have more familiarity with scores, grades, and other summative evaluations. Such evaluations may seem more objective and criteria-based to students, and thus more amenable to automation. As a result, initial expectations about both scoring and feedback might have been anchored toward scoring capabilities. By contrast, formative assessment is less common and may be viewed as more subjective and personal. Students may have had a weaker, less confident foundation for judging whether a computer could offer such feedback. They might have adopted a "wait and see" approach to evaluating automated feedback that weighted direct experience much more strongly than any initial expectations. In future work, it may be valuable to qualitatively probe students' underlying reasoning and confidence in their ratings to better understand the prior knowledge, beliefs, or certainty that guide such judgments.

Overall, these findings inspire a qualification to our answer to the first research question. Fortunately, AWE developers may not be entirely at the mercy of their detractors! Although instructors or peers might present the software in a negative light, a well-designed system that seems to give accurate scores with valid, useful, and empowering formative recommendations might override initial doubts. In other words, pedagogical principles of effective feedback still apply to the design of automated feedback, and might be crucial to winning over skeptical student, teacher, and policymaker audiences. We know that feedback varies in effectiveness based on content and contextual factors (Hattie &

Timperley, 2007; Kluger & DeNisi, 1998; Shute, 2008). For instance, summative and formative feedback are better when they support strategic processing, metacognition, self-regulation (Hattie & Timperley, 2007; Kluger & DeNisi, 1998), and feedback effectiveness can also depend on timing, specificity, and students' achievement level. (Shute, 2008).

In future research, AWE developers may wish to expand beyond comparisons of human and automated scoring (e.g., Bridgeman et al., 2012) to consider the intersection of feedback design and AWE constraints (e.g., Roscoe, Varner, Crossley, & McNamara, 2013). The limitations of NLP will always place restrictions on what features of writing can be reliably targeted for feedback. Nonetheless, there may be creative ways to sidestep these limitations in order to provide feedback on issues other than word count, spelling, organization, semantic similarity to source documents, and so on. Ultimately, and most importantly, it behooves us to make sure that anything we *can* comment upon via automated feedback is done so in accord with best practices in feedback design.

4.3. Effects of perceptions on writing and future intentions

Our final analyses explored how students' perceptions were related to their writing performance and behaviors, and to their future intentions regarding the software (RQ3). As discussed above, a key motivation for studying user perceptions is that those perceptions can influence how a technology is adopted and used (Ertmer et al., 2012; Vinkatesh & Davis, 2000). In the case of AWE, this refers to how the software influences students' essay quality and revising, and whether students want to keep using it. Evaluations of AWE systems have found that large proportions of students do not revise (Wilson et al., 2014), which suggests that many students either disengage from the software or the software is not implemented in a manner conducive to revising (e.g., schools using AWE to support benchmarking rather than the development of writing skills). If students do not or cannot take full advantage of writing practice opportunities and feedback, this may undermine the effectiveness of AWE in the classroom (see Stevenson & Phakiti, 2013).

Students' expectations, immediate perceptions, and final impressions of W-Pal had little to no relation to their revising behaviors. Data show that students did revise by implementing additions, deletions, and substitutions at both the word and sentence level. These revisions also included superficial revisions that preserved the meaning of the text and substantive revisions that transformed the text. Such edits resulted in moderate gains in essay quality, which were correlated with additions and substantive edits. However, although students were able to use W-Pal feedback to revise and improve their writing, these behaviors were not influenced by their perceptions. Students revised regardless of whether they held positive or negative expectations or perceptions of the scoring and feedback.

These results parallel findings from research on perceptions of human-delivered feedback. For example, Kaufman and Schunn (2011) examined undergraduate's perceptions of peer feedback and their subsequent revisions. Students rated the usefulness, positivity, validity, reliability, and fairness of the feedback they received from peers. Essay revisions were coded as simple changes involving three or fewer consecutive words, or as complex changes involving four or more consecutive words. Complex changes were positively correlated with gains in peer-assigned scores across drafts but simple changes were not. Most importantly, perceptions of feedback quality were unrelated to either type of revision. Similar results were observed by Strijbos et al. (2010), who examined graduate students' perceptions of writing feedback in relation to their ability to revise a text. Students received a pre-prepared text with feedback and were asked to rate the feedback as if they had received it on their own writing. Students were then instructed to revise the text. Feedback perceptions were unrelated to whether errors were correctly identified, explained, or resolved.

Although current results and prior studies suggest that feedback perceptions may not directly hinder writing performance or behaviors, our findings build on this pattern in a crucial way. When we probed students' willingness to use W-Pal again in the future, or to recommend it to a friend who needed assistance, the majority of students responded affirmatively. Positive and negative responses, however, were predictable based on students' perceptual changes in feedback quality. When students felt that the feedback they received was better than expected (e.g., more accurate, more relevant, or more useful), they demonstrated a very high chance of willingness to use W-Pal to improve their own writing or suggest it to a friend. However, disappointment in the system resulted in the decision to not use or recommend the system.

One interpretation of these results is that, in the moment, students can find value even in flawed AWE feedback (e.g., Grimes & Warschauer, 2010). However, when probed about future intentions, students' positive and negative perceptions become much more important. There are consequences for students' negative perceptions and experiences of AWE, but these consequences may not manifest immediately. Overall, presentation may set the stage, perhaps in subtle ways, to influence expectations and reactions, which in turn shape final impressions about AWE. These experiences then affect future intentions.

4.4. Limitations and future research

There are several limitations to the current study that could be addressed in future research that collects more data, over a longer period of time, and in authentic classrooms.

First, although we attempted to probe students' perceptions of feedback accuracy, relevance, and usefulness separately, individual ratings tended to load on a single dimension. These results corroborate prior research on feedback perceptions suggesting that students make holistic rather than nuanced judgments about feedback quality (Strijbos et al., 2010). Thus, for automated feedback to be appreciated by student users, that feedback may have to satisfy multiple constraints. For example, if the feedback communicates “great strategies” (i.e., high utility) but seems disconnected from specific elements of students' writing (i.e., low relevance), the feedback could still be judged negatively or rejected. Consequently, one approach for future work may be to focus parsimoniously on only holistic judgments of feedback. Alternatively, a lack of nuance in students' judgments may represent lack of knowledge or skills related to writing assessment. Teaching students to implement detailed assessment rubrics might not only improve writing proficiency (Panadero & Jonsson, 2013), but could hypothetically encourage more fine-grained evaluations of automated feedback received.

Another potential limitation is that this study did not include a direct assessment students' uptake of specific recommendations in W-Pal feedback. Thus, we cannot know the extent to which feedback recommendations were adopted or rejected, and we may be overlooking subtle effects of perceptions on revising. Importantly, students could receive multiple recommendations on multiple topics, and their reasons for following or ignoring each suggestion could be diverse. Students might have implemented a recommendation because they viewed it as important or, alternatively, because it seemed to require the least effort. Similarly, when students did not implement a recommendation, we could not know whether that action was intentional (e.g., they disagreed with it) or accidental (e.g., they did not read the feedback carefully).

The purpose of the current study was to explore students' perceptions of AWE, sources of these perceptions, and potential consequences for writing and user intentions, but not to conduct extensive investigations of writing composition processes or human-computer interactions. However, in future research, one valuable approach may be to use cued retrospective reports (see Van Gog, Paas, van Merrienboer, & Witte, 2005) to probe students' decisions and underlying reasoning regarding feedback and revisions. Screen capture technologies can record the interface as students write and revise, and the resulting videos could be reviewed by the students to elicit their commentary. For instance, whenever a revision event occurs, students could be asked to reflect on their goals or motivations for editing their essay. Similarly, students could review their automated feedback reports and be prompted to explain their responses and decisions regarding the presented messages. These qualitative self-report data might provide a deeper understanding of students' nuanced perceptions of automated scoring and feedback, and a better understanding of the how these perceptions might influence decision making while using AWE software.

A third limitation is that this study comprised only a single cycle of writing and revising. All writing, revising, and survey activities occurred in a 1-h session. In a more authentic writing task or classroom environment, students would likely spend more time working within the AWE system as they completed multiple assignments over weeks, months, or the school year. Thus, students' typical AWE interactions are not one-shot experiences as in the current study—they are iterative and longitudinal interactions. In addition, students working in real classrooms have different stakes for their writing than students participating in a research study. Our research participants may not have taken the writing task as seriously. However, given that students did spend much of their allotted time on-task, and made real efforts to revise, it suggests that they were engaged. The selected writing prompt (i.e., psychology in the media) was related to their introductory psychology course.

In future work, a longitudinal design that examines patterns of writing proficiency, revising behaviors, and AWE perceptions over time may be very powerful for understanding how perceptions influence AWE efficacy and use. Current results found that initial expectations and immediate perceptions influenced final perceptions, which in turn guided future intentions with W-Pal. Students who found the system to be unhelpful, or less helpful than expected, did not wish to continue using it. An open question is what happens when a student who "hates" or "distrusts" an AWE system is forced to use it to write and revise multiple essays. Although a series of positive interactions might eventually "win them over," it is also plausible that initial negative impressions inspire a further downward spiral across future iterations. How do perceptions and intentions formed after one instance of using AWE software inform expectations for the next encounter? Moreover, how malleable or entrenched are students' beliefs, and over what time course do such perceptions become stable or durable?

A longitudinal design that includes more student writers, essays, and perceptions would also enable more complex consideration of mediation, moderation, or dynamic relationships among variables. We observed hints that expectations and perceptions at different stages may have both direct and indirect effects on each other, although we focused on direct effects in this study. One hypothesis is that the relative importance and predictive value of expectations versus immediate perceptions may change over time. When students first begin to use an AWE system, they may interact with the system based on loosely-held expectations and biases stemming from instructors' viewpoints, word-of-mouth, or similar second-hand sources. During early uses of the system, students' immediate reactions to the software might quickly supplant initial, vague expectations. When students receive useful feedback and experience growth in their writing ability, they may feel more favorable toward the system. In contrast, repeated disappointment due to "unfair scores" or "confusing feedback" may inspire frustration. Over multiple interactions with an AWE system, these immediate perceptions may converge into stable beliefs or "grounded expectations" about AWE that are less affected by any single encounter.

5.0. Concluding remarks

The current research adopted what might be considered a "user experience" approach that emphasized user (i.e., student) technology perceptions and their relationship with outcomes. Naturally, a user experience approach is not the only valid means for improving AWE. We view this research as only one aspect of a multifaceted strategy that can and should also include "software engineering" and "pedagogical" approaches. From a software standpoint, it remains necessary to develop new and better algorithms for detecting key features of writing. Common concerns about AWE have focused on what the software cannot do (see Deane, 2013)—computers cannot understand humor, logic, types of persuasion, and so on (Anson et al., 2013). Thus, one way to advance AWE capabilities is to push their computational boundaries. Indeed, as just one example, scholars are currently working on NLP approaches to humor detection that could be applied to AWE systems (e.g., Skalicky, Berger, Crossley, & McNamara, 2016). Within a corpus of over 300 undergraduate student essays, Skalicky and colleagues were recently able to use NLP-detected linguistic features of text to account for 17.5% of the variance in human judgments of humor in writing.

Similarly, we may also need to improve the instructional and pedagogical foundation of AWE systems, such as crafting better and more personalized formative feedback messages, or incorporating explicit strategy instruction or motivating games (Allen et al., 2014; Roscoe et al., 2014). A wealth of research already exists regarding effective formative feedback (Parr & Timperley, 2010; Shute, 2008) and writing instruction (Graham & Perin, 2007; Graham, MacArthur, & Fitzgerald, 2013; Mason, Harris, & Graham, 2011) that might be tapped more deeply to inform AWE design. However, discussion of sound instructional principles seems to be neglected in the AWE literature relative to issues of accuracy and statistical modeling.

Acknowledgements

This research was supported in part by a grant from the Institute of Education Sciences (IES R305A120707). Opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the Institute of Education Sciences. The authors would like to thank Danielle McNamara, Laura Allen, Matthew Jacovina, and Jianmin Dai for their input and support.

Appendix

Students in each condition received a similar introduction to the W-Pal software, but the relative strengths or weaknesses of scoring and feedback were manipulated. The following excerpts provide the scripts as presented to students.

Weak Scoring/Weak Feedback Condition Script

"Okay! So let me tell you more about the software you'll be using today, which is called Writing Pal. We're currently in the

process of improving how it automatically grades students' essays and improving how it gives them feedback on their writing.

One purpose of this study is to help us identify and fix issues in the scoring system. It is not easy for computers to automatically grade an essay. It takes a lot of testing to improve the accuracy of the scoring. In our research, we use a 6-point essay grading scale. Compared to the scores a human expert would give, the computer scores can sometimes be off by several points.

At the same time, we're also trying to identify and fix issues in the feedback system. It is very challenging to program computer software to give writing feedback that is accurate, on topic, and helpful to the writer. Our goal is to make the automated feedback more understandable and specific for individual writers, but that is definitely a work-in-progress.

So, basically, in this study you will be helping us test and improve our software."

Strong Scoring/Weak Feedback Condition Script

"Okay! So let me tell you more about the software you'll be using today, which is called Writing Pal. The system has already been designed to automatically grade students' essays with a high level of accuracy, and we're currently in the process of improving how it gives them feedback on their writing.

For over five years, we have been developing the scoring system based on input from student users and from experts in writing, linguistics, psychology, and computer science. In our research, we use a 6-point grading scale. Compared to the scores that a human expert would give, the computer scores are very similar or even an exact match.

At the same time, we're also trying to identify and fix issues in the feedback system. It is very challenging to program computer software to give writing feedback that is accurate, on topic, and helpful to the writer. Our goal is to make the automated feedback more understandable and specific for individual writers, but that is definitely a work-in-progress.

So, basically, in this study you will be helping us test and expand our software."

Weak Scoring/Strong Feedback Condition Script

"Okay! So let me tell you more about the software you'll be using today, which is called Writing Pal. The system has already been designed to automatically give valid and meaningful feedback on students' writing. We're currently in the process of improving how it automatically grades students' essays.

One purpose of this study is to help us identify and fix issues in the scoring system. It is not easy for computers to automatically grade an essay. It takes a lot of testing to improve the accuracy of the scoring. In our research, we use a 6-point essay grading scale. Compared to the scores a human expert would give, the computer scores can sometimes be off by several points.

However, for over five years, we have been developing the feedback system based on input from experts in writing, linguistics, psychology, and computer science. We have designed the computer software to give feedback that is accurate, on topic, and helpful to the writer. We have also made sure the feedback is understandable and specific for individual writers. In fact, our research shows that the feedback can help students raise their score.

So, basically, in this study you will be helping us test and expand our software."

Strong Scoring/Strong Feedback Condition Script

"Okay! So let me tell you more about the software you'll be using today, which is called Writing Pal. The system has already been designed to automatically grade students' essays with a high level of accuracy and give valid and meaningful feedback on students' writing. We're currently in the process of testing the system.

For over five years, we have been developing the scoring system based on input from student users and from experts in writing, linguistics, psychology, and computer science. In our research, we use a 6-point grading scale. Compared to the scores that a human expert would give, the computer scores are very similar or even an exact match.

We have designed the computer software to give feedback that is accurate, on topic, and helpful to the writer. We have also made sure the feedback is understandable and specific for individual writers. In fact, our research shows that the feedback can help students raise their score.

So, basically, in this study you will be helping us test our software."

References

Allen, L. K., Crossley, S. A., Snow, E. L., & McNamara, D. S. (2014). L2 writing practice: Game enjoyment as a key to engagement. *Language Learning and Technology, 18*, 124–150.

Anson, C., Filkins, S., Hicks, T., O'Neill, P., Pierce, K. M., & Winn, M. (2013). *NCTE position statement on machine scoring: Machine scoring fails the test*. National Council of Teachers of English. Retrieved from http://www.ncte.org.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*, 27–40.

Bridwell, L. S. (1980). Revising strategies in twelfth grade students' transactional writing. *Research in the Teaching of English, 14*, 197–222.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*, 100–108.

Crawford, L., Lloyd, S., & Knoth, K. (2008). Analysis of student revisions on a state writing test. *Assessment for Effective Intervention, 33*, 108–119.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*, 1227–1237.

Crossley, S. A., Roscoe, R. D., & McNamara, D. S. (2013). Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In C. Boonthum-Denecke, & G. M. Youngblood (Eds.), *Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference* (pp. 208–2013). AAAI Press.

Crossley, S. A., Weston, J. L., Sullivan, S. T. M., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*, 282–311.

Curran, M. J., Draus, P., Maruschock, G., & Maier, T. (2013). Student perceptions of Project Essay Grade (PEG) software. *Issues in Information Systems, 14*, 89–98.

Curran, M. J., Draus, P., Maruschock, G., & Maier, T. (2014). Faculty perceptions of Project essay Grade (PEG) software. *Issues in Information Systems, 15*, 71–80.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7–24.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning and Assessment, 5*. Retrieved from http://www.jtla.org.

Ertmer, P. A. (1999). Addressing first- and second-order barriers to change:

Strategies for technology integration. *Educational Technology Research and Design, 47*, 47–61.
Ertmer, P. A., Ottenbreit-Leftwich, A. T., Sadik, O., Sendurur, E., & Sendurur, P. (2012). Teacher beliefs and technology integration practices: A critical relationship. *Computers and Education, 59*, 423–435.
Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication, 32*, 400–414.
Fellbaum, C. (1998). *WordNet*. Blackwell Publishing Ltd.
Foltz, P. (2007). Discourse coherence and LSA. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 167–184). Lawrence Erlbaum Associates.
Foltz, P., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68–88). Routledge.
Graham, S., MacArthur, C. A., & Fitzgerald, J. (2013). *Best practices in writing instruction*. New York, NY: Guilford Press.
Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445–476.
Grimes, D., & Warschauer, M. (2010). Utility in a fallible took: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*, 4–43.
Hattie, J., & Timperley, H. S. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.
Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*, 369–388.
Hearst, M. A. (2000). The debate on automated essay grading. *Intelligent Systems and their Applications, IEEE, 15*, 22–37.
Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science, 39*, 387–406.
Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin and Review, 14*, 237–242.
Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research, 42*, 173–196.
Kluger, A. N., & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science, 7*, 67–72.
Koehler, M. J., & Mishra, P. (2009). What is technological pedagogical content knowledge? *Contemporary Issues in Technology and Teacher Education, 9*, 60–70.
Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*, 757–786.
Lai, Y. (2010). Which do students prefer to evaluate their essays: Peers or computer program? *British Journal of Educational Technology, 41*, 432–454.
Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *27*, 1–18.
Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied, 5*, 319–333.
MacArthur, C. A., Philippakos, Z. A., & Ianetta, M. (2015). Self-regulated strategy instruction in college developmental writing. *Journal of Educational Psychology, 107*, 855–867.
Mason, L. H., Harris, K. R., & Graham, S. (2011). Self-Regulated Strategy Development for students with writing difficulties. *Theory into Practice, 50*, 20–27.
McGarrell, H., & Verbeem, J. (2007). Motivating revision of drafts through formative feedback. *ELT Journal, 61*, 228–236.
McNamara, D. S., Crossley, S. A., & Roscoe, R. D. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods, 45*, 499–515.
McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35–59.
McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.
National Governors Association for Best Practices. (2010). *Common core state Standards: English language arts*. Washington, DC: National Governors Association for Best Practices, Council of Chief State School Officers.
Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144.
Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*, 68–85.
Roscoe, R. D., Allen, L., Weston, J., Crossley, S., & McNamara, D. S. (2014). The writing pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*, 39–59.
Roscoe, R. D., & McNamara, D. S. (2013). Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105*, 1010–1025.
Roscoe, R. D., Jacovina, M. E., Harry, D., Russell, D. G., & McNamara, D. S. (2015). Partial verbal redundancy in multimedia presentations for writing strategy instruction. *Applied Cognitive Psychology, 29*, 669–679.
Roscoe, R. D., Snow, E. L., Allen, L. K., & McNamara, D. S. (2015). Automated detection of essay revision patterns: Applications for intelligent feedback in a writing tutor. *Technology, Instructional, Cognition, and Learning, 10*, 59–79.
Roscoe, R. D., Varner (Allen), L. K., Crossley, S. A., & McNamara, D. S. (2013). Developing pedagogically-guided algorithms for intelligent writing feedback. *International Journal of Learning Technology, 8*, 362–381.
Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76.
Shermis, M. D., & Burstein, J. C. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189.
Skalicky, S., Berger, C. M., Crossley, S. A., & McNamara, D. S. (2016). Linguistic features of humor in academic writing. *Advanced in Language and Literature Studies, 7*, 248–259.
Sommers, N. (1982). Responding to student writing. *College Composition and Communication, 33*, 148–156.
Stevenson, M., & Phakiti, A. (2013). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.
Strijbos, J., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*, 291–303.
Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24–54.
Van Gog, T., Paas, F., van Merriënboer, J. J. G., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied, 11*, 237–244.
Vinkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: For longitudinal studies. *Management Science, 46*, 186–204.
Voogt, J., Fisser, P., Pareja Roblin, N., Tondeur, J., & van Braak, J. (2013). Technological pedagogical content knowledge: A review of the literature. *Journal of Computer Assisted Learning, 29*, 109–121.
Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*, 22–36.
Webb, N. M., Nemer, K. M., & Ing, M. (2006). Small-group reflections: Parallels between teacher discourse and student behavior in peer-directed groups. *Journal of the Learning Sciences, 15*, 63–119.
Wilson, J., Olinghouse, N. G., & Andrada, G. N. (2014). Does automated feedback improve writing quality? *Learning Disabilities: A Contemporary Journal, 12*, 93–118.
Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When student belief that personal characteristics can be developed. *Educational Psychologist, 47*, 302–314.